Refined PAC-Bayes Bounds for Offline Bandits

Amaury Gouverneur, Tobias J. Oechtering, and Mikael Skoglund
Division of Information Science and Engineering (ISE)
KTH Royal Institute of Technology, Stockholm
{amauryg,oech,skoglund}@kth.se

***Abstract*—In this paper, we present refined probabilistic bounds on empirical reward estimates for off-policy learning in bandit problems. We build on the PAC-Bayesian bounds from [1]–[4] and improve on their results using a new parameter optimization approach introduced by [5]. This technique is based on a discretization of the space of possible events to optimize the "in probability" parameter, λ. We provide two parameter-free PAC-Bayes bounds, one based on Hoeffding-Azuma's inequality and the other based on Bernstein's inequality. We prove that our bounds are *almost optimal* as they recover the same rate as would be obtained by setting the "in probability" parameter after the realization of the data.** [1]

I. Introduction

In bandit problems, an agent interacts sequentially for T times with an unknown environment. At each time $t \in \{1, \dots, T\}$, the agent select an action $a_t \in \mathcal{A}$ according to a decision strategy π, a policy, expressed as a probability distribution over the possible actions. Based on the chosen action, the environment produces a reward. By learning from the observed action-reward pairs, the agent aims to identify actions that maximize the long-term cumulative reward. This decision-making framework has applications in diverse domains, including healthcare, finance, recommender systems, and telecommunications (see [6, 7] for a survey).

In the *offline* bandit setting, the agent is provided with a dataset h^t consisting of previously observed action-reward pairs, $\{a_n, r_n\}_{n=1}^t$. Based on this dataset, the agent must select a fixed policy that yields high rewards *in expectation*. Since the reward distribution is unknown, the agent relies on an *empirical estimate*, $\widehat{\mathscr{R}}(\pi, h^t)$, to evaluate the expected reward associated with the policy π. A important question to guide the policy search is whether the empirical estimate $\widehat{\mathscr{R}}(\pi, h^t)$ is close to the true expected reward $\mathscr{R}(\pi)$.

Probably approximately correct (PAC) theory provides answers to this question by offering "in probability" guarantees on the difference between the empirical estimate and the true expectation. Traditionally, these bounds have been developed for supervised learning problems and depend on the complexity of the hypothesis space, which can be measured using quantities like the Vapnik–Chervonenkis dimension or Rademacher complexity, as discussed in [8]. PAC-Bayesian bounds generalize PAC guarantees and account for the dependence of the learned hypothesis on the observed dataset [9]–[13]. This dependence is typically measured using the relative entropy between a prior hypothesis and the learned hypothesis.

Initially focused on supervised learning problems where the dataset consists of independent and identically distributed samples, the PAC-Bayes framework was later extended to bandit problems [1]. This extension leverage martingale properties to address the dependencies inherent in sequential decision problems. The resulting bounds have found practical applications, as they motivated the introduction of relative entropy regularization methods. The idea is to search for a policy π that maximizes the expected reward estimate while imposing a penalty on the relative entropy between π and a prior policy μ. Examples of such methods include Relative Entropy Policy Search [14], Trust Region Policy Optimization [15], and Proximal Policy Optimization [16].

One problematic of the PAC-Bayes bounds derived in [1] and follow-up works [2, 3, 17] is that they depend on a parameter $\lambda > 0$ that must be set before observing the data and cannot be optimized based on the data. [17] attempts to circumvent this difficulty deriving a bound that holds simultaneously for a grid of parameters, but their approach fells short to obtain a parameter-free bound with optimal rate.

In this work, we improve on these previous results and derive parameter-free PAC-Bayes bounds, achieving the optimal rate. We apply an optimization technique for the "in probability" parameter λ that works by discretizing the *space of possible events* for the bounds and optimizing the parameter λ *conditioned on the event* before applying a union bound [5]. We provide two optimized PAC-Bayes bounds: one based on the Hoeffding-Azuma inequality and one on the Bernstein inequality.

The rest of the paper is organized as follows:

- Section II introduces the notation, presents the multi-armed bandit problem and contextual bandit problem, and discusses the online and offline learning settings.
- Section III explains the importance sampling estimate before proving different PAC-Bayes bounds.
- Finally, our conclusion is presented in Section IV.

[1] This work was supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

II. Preliminaries

A. General Notation

Random variables X are written in capital letters, their realizations x in lowercase letters, their outcome space in calligraphic letters $\mathcal{X}$, and their distribution is written as $\mathbb{P}_X$. The density of a random variable X with respect to a measure μ is written as $f_X := \frac{d\mathbb{P}_X}{d\mu}$. When two (or more) random variables X, Y are considered, the conditional distribution of Y given X is written as $\mathbb{P}_{Y|X}$, and the notation is abused to write their joint distribution as $\mathbb{P}_X \mathbb{P}_{Y|X}$.

We use the underscore notation X_t to represent a random variable at time $t = 1, \ldots, T$ and the exponent notation X^t to denote a sequence of random variables $X^t \equiv (X_1, \ldots, X_t)$ for $t = 2, \ldots, T$. For consistency we let $X^1 \equiv X_1$.

We use the notation $d\mathbb{P}/d\mathbb{Q}$ for the Radon-Nikodym derivative. The relative entropy between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ is defined as $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) := \int \log\left(\frac{d\mathbb{P}}{d\mathbb{Q}}\right) d\mathbb{P}$ if $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ and $D_{\text{KL}}(\mathbb{P} \parallel \mathbb{Q}) \to \infty$ otherwise.

B. Multi-Armed Bandits

A *multi-armed bandit* is a sequential decision problem where, at each time step $t \in \{1, \ldots, T\}$, an agent chooses an action $A_t \in \mathcal{A}$ according to a decision policy $\pi_t \in \Pi$, expressed as a probability distribution over the possible actions; that is $\pi_t(a)$ expresses the probability that agent takes action a at time t. The agent then receives a positive random reward $R_t \in [0, 1]$ distributed according to some fixed but unknown reward distribution $\mathbb{P}_{R_t|A_t}$. As the reward distribution depends on the chosen action, it can be written as $R_t = R(A_t)$ for some random function R. We use the notation $\overline{r}(A_t) := \mathbb{E}[R(A_t)]$ to denote the expected reward associated with action A_t. The data is collected in a history $H^t = H^{t-1} \cup \{A_t, R_t\}$, where H^{t-1} contains all action-reward pairs observed prior to t.

We introduce the notation $\mathscr{R}(\pi)$ to denote the expected reward obtained by taking actions according to a policy π, with the expectation taken both over the randomness of the action selection and the reward distribution:

$$\mathscr{R}(\pi) := \mathbb{E}_{A \sim \pi(\cdot)}[\overline{r}(A)]. \tag{1}$$

Example 1 (Drug dosage): In clinical trials for a new drug, selecting the dosage is critical. Each action $A_t \in \mathcal{A}$ represents a different dosage level, and the reward R_t corresponds to the patient's response to the dosage, such as an improvement in health metrics or the absence of adverse side effects.

C. Contextual Bandits

In a *contextual bandit* problem, at each time step $t \in \{1, \ldots, T\}$, the agent first observes a context $X_t \in \mathcal{X}$ sampled from a fixed but unknown distribution $\mathbb{P}_X$. Based on the context, the agent selects an action $A_t \in \mathcal{A}$ according to a decision policy. Here, the decision policy, $\pi_t \in \Pi$, represents a probability distribution over the action given the observed context. That is, $\pi_t(a, x)$ gives the probability that the agent takes action a given that he observed context x. The agent then receives a reward $R_t \in \mathbb{R}$ from a fixed but unknown distribution that depends on the context and the action taken, $\mathbb{P}_{R_t|A_t,X_t}$. The reward can be written $R_t = R(A_t, X_t)$ for some random function R. We use the notation $\overline{r}(A_t, X_t) := \mathbb{E}[R(A_t, X_t)]$ to denote the expected reward for action A_t and context X_t. The data is collected in a history $H^t = H^{t-1} \cup H_t$, where $H_t = \{A_t, X_t, R_t\}$. Accordingly, we introduce the notation $\mathscr{R}(\pi)$ to denote the expected reward obtained by taking actions according to a policy π:

$$\mathscr{R}(\pi) := \mathbb{E}_{X \sim \mathbb{P}_X, A \sim \pi(\cdot, X)}[\overline{r}(A, X)]. \tag{2}$$

Example 2 (Online advertising): In online advertising, advertisers select which ad to show to a user based on user data or context $X_t \in \mathcal{X}$, such as demographics, browsing history, or location. Each action $A_t \in \mathcal{A}$ corresponds to a different advertisement, and the reward R_t represents an outcome, such as whether the user clicks on the ad or the revenue generated.

D. Online and Offline Settings

In the *online* setting, the goal of the agent is to find a *sequence* of policies, $\pi_{1:T} = \{\pi_1, \pi_2, \ldots, \pi_T\}$ that maximizes the expected *cumulative reward*, $\sum_{t=1}^{T} \mathscr{R}(\pi_t)$, or equivalently that minimizes the *cumulative regret*, defined as the expected difference between the best cumulative reward and the cumulative reward obtained following the policies:

$$\operatorname{reg}(\pi_{1:T}) := \sum_{t=1}^{T} \mathscr{R}(\pi^\star) - \mathscr{R}(\pi_t),$$

where $\pi^\star \in \operatorname{argmax}_{\pi \in \Pi} \mathscr{R}(\pi)$ is an optimal policy. To ensure that such a policy exists, we make the technical assumption that the set of policies Π is compact. In this setting, each policy π_t is chosen based on the observed data h^t, and the agent faces the *exploration-exploitation* dilemma: should it try less explored actions to learn more about the reward distribution or exploit the best-performing ones based on the information gathered.

In the *offline* setting, also known as batch learning [18], the agent is given a fixed set of logged data $h^t = \{a_n, r_n\}_{n=1}^{t}$, where each action A_n was sampled from a known logging policy π_n, also known as behavior policies, and each reward r_n obtained from the unknown reward distribution $\mathbb{P}_{R|A=a_n}$. Note that if the logging policies π_n are the same for all $n \in \{1, \ldots, t\}$ then the collected data consists of i.i.d. samples. Given the dataset, h^t, the agent's goal in the offline setting is to select a single *target* policy $\pi \in \Pi$ that improves on the logged policies and will yield the highest expected reward.

The choice to perform online or offline learning typically depends on the nature of the application. One could argue that Example 1 is better suited for offline learning than Example 2.

III. PAC-Bayes Bounds

In both the online and offline setting, maximizing directly the expected reward $\mathscr{R}(\pi)$ over the policies $\pi \in \Pi$ is not possible as the reward distribution $\mathbb{P}_R$ is unknown. To guide the policy search of a policy, we have to use *empirical estimate* of the expected reward $\mathscr{R}(\pi)$ based on the observed data H^t.

One commonly used estimate is the importance sampling (IS) estimate [19]. For the multi-armed bandit setting, the importance sampling estimate for the expected reward associated with action $a \in \mathcal{A}$ is given by:

$$\hat{r}^{\text{IS}}(a, h^t) \coloneqq \frac{1}{t} \sum_{n=1}^{t} \frac{\mathbb{1}_a\{a_n\}}{\pi_n(a_n)} r_n, \tag{3}$$

where the indicator function $\mathbb{1}_a\{a_n\}$ returns 1 if $a_n = a$ and 0 otherwise. We can verify that $\hat{r}^{\text{IS}}(a, h^t)$ is an unbiased estimator of $\overline{r}(a)$ as taking the expectation of the dataset H^t gives $\mathbb{E}[\hat{r}^{\text{IS}}(a, H^t)] = \overline{r}(a)$.

For contextual bandits, let $n(x, h^t) = \sum_{n=1}^{t} \mathbb{1}_x\{x_n\}$ denote the number of times that the context $x \in \mathcal{X}$ appeared in the dataset h^t. The importance sampling estimate of the expected reward associated with action $a \in \mathcal{A}$ and context $x \in h^t$ is given by

$$\hat{r}^{\text{IS}}(a, x, h^t) \coloneqq \frac{1}{n(x, h^t)} \sum_{n=1}^{t} \frac{\mathbb{1}_{(a,x)}\{(a_n, x_n)\}}{\pi_n(a_n, x_n)} r_n. \tag{4}$$

When using these empirical estimates, one important question is whether the policy estimate $\widehat{\mathscr{R}}^{\text{IS}}(\pi, h^t) = \mathbb{E}_{A\sim\pi}[\hat{r}^{\text{IS}}(A, h^t)]$ is close to the actual policy value $\mathscr{R}(\pi)$. PAC-Bayes theory provides answers to that question as it offers in probability guarantees on the difference between $\mathscr{R}(\pi)$ and $\widehat{\mathscr{R}}^{\text{IS}}(\pi, h^t)$.

A. Bandit Problems and Martingales

The bandit setting introduces challenges not encountered in typical PAC-Bayesian learning settings. In bandit problems, the data H^t is often neither independent nor identically distributed, as it depends on the sequential actions which, in turn, depend on the previously observed data. These challenges can be addressed by working with *martingales*.

Definition 1 (Martingale): A sequence of random variables M^t is called a martingale with respect to X^t if the following conditions hold for every $n \in \{1, \dots, t\}$:

1) M_n is completely determined by $X_1, \dots, X_n$,
2) $\mathbb{E}[|M_n|] < \infty$,
3) $\mathbb{E}[M_n \mid X_1, \dots, X_{n-1}] = M_{n-1}$.

The sequence of consecutive differences in a martingale sequence $Z_n = M_n - M_{n-1}$ is referred to as a *martingale difference sequence* and satisfies $\mathbb{E}[Z_n \mid X_1, \dots, X_{n-1}] = 0$.

One reason to work with the importance sampling estimate is that it naturally gives rise to martingale sequences. For instance, for any action $a \in \mathcal{A}$, the sequence of random variables

$$M_n^{IS}(a) = n\big(\hat{r}^{\text{IS}}(a, H^n) - \overline{r}(a)\big), \tag{5}$$

forms a martingale sequence with respect to H^t and

$$Z_n^{IS}(a) = \frac{\mathbb{1}_a\{A_n\}}{\pi_n(A_n)} R_n - \overline{r}(a), \tag{6}$$

forms its corresponding martingale difference sequence. A proof of that statement is provided in [20, Lemma B.1].

B. PAC-Bayes Bound for $\widehat{\mathscr{R}}^{\text{IS}}(\pi, H^t)$

Deriving PAC-Bayes bounds for martingales can be done with two main tools: Donsker-Varadhan's variational formula for relative entropy [21] and a martingale concentration inequality, such as Hoeffding-Azuma inequality [22, 23] or Bernstein inequality [24, Proposition 2.10]. We recall these classical results below.

Lemma 1 (Donsker-Varadhan's variational formula): For any measurable, bounded function $h : \mathcal{A} \to \mathbb{R}$ and any probability distribution on $\mathcal{A}$, $\mu \in \Pi$, such that $\mathbb{E}_{A\sim\mu}[e^{h(A)}] < \infty$, we have

$$\log \mathbb{E}_{A\sim\mu}[e^{h(A)}] = \sup_{\pi\in\Pi} (\mathbb{E}_{A\sim\pi}[h(A)] - D_{\text{KL}}(\pi \,\|\, \mu)).$$

Lemma 2 (Hoeffding-Azuma's inequality): Let Z^t be a martingale difference sequence, such that for all $n \in \{1, \dots, t\}$ $Z_n \in [a_n, b_n]$ and let $M_t = \sum_{n=1}^{t} Z_n$ be the corresponding martingale. Then, for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[e^{\lambda M_t}] \le e^{\frac{\lambda^2 \sum_{n=1}^{t}(b_n - a_n)^2}{8}}.$$

Lemma 3 (Bernstein's inequality):
Let Z^t be a martingale difference sequence such that $|Z_n| \le [a, b]$ for all $n \in \{1, \dots, t\}$ with probability 1. Let $M_t = \sum_{n=1}^{t} Z_n$ be a corresponding martingale and $V_t = \sum_{n=1}^{t} \mathbb{E}[Z_n^2 \mid Z_1, \dots, Z_{n-1}]$ be the cumulative variance of this martingale. Then for any fixed $\lambda \in [0, \frac{1}{b-a}]$, we have

$$\mathbb{E}[e^{\lambda M_t}] \le \mathbb{E}[e^{\lambda^2 V_t (e-2)}].$$

The first PAC-Bayes bound we present was introduced in [2], and as in [4, Theorem 5], we provide an alternative proof based on Donsker-Varadhan variational formal and Hoeffding-Azuma's inequality.

Proposition 1 (Hoeffding PAC-Bayes bound for $\hat{r}^{\text{IS}}$): Let $\mu \in \Pi$ be any prior policy independent of H^t and let ε be a uniform lower bound on $\{\pi_n(a)\}_{n=1}^{t}$ for all $a \in \mathcal{A}$ (resp. on $\{\pi_n(a, x)\}_{n=1}^{t}$ in the contextual bandits setting) . Then, for any $\lambda > 0$, and any $\beta \in (0, 1)$, with probability no smaller than $1 - \beta$ (over the sampling of H^t)

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\text{IS}}(\pi, H^t)| \le \frac{\lambda}{8t\varepsilon^2} + \frac{D_{\text{KL}}(\pi \,\|\, \mu) + \ln\frac{2}{\beta}}{\lambda},$$

holds simultaneously for all policy $\pi \in \Pi$ depending on H^t.

Proof: We start by fixing $a \in \mathcal{A}$ and observe that the elements of the martingale difference sequence $\{Z_n^{IS}(a)\}_{n=1}^t$, defined in eq. (6), are in the interval $[-\overline{r}(a), \frac{1}{\varepsilon} - \overline{r}(a)]$. Applying Lemma 2 on the corresponding martingale sequence $\{n\big(\hat{r}^{\mathrm{IS}}(a, H^n) - \overline{r}(a)\big)\}_{n=1}^t$, for any $l > 0$, we have that

$$\mathbb{E}_{H^t}\big[e^{lt\left(\hat{r}^{\mathrm{IS}}(a,H^t) - \overline{r}(a)\right)}\big] \leq e^{\frac{l^2 t}{8\varepsilon^2}}.$$

We set $\lambda = l/t$ and integrate over $a \in \mathcal{A}$ with respect to the distribution μ, which gives

$$\mathbb{E}_{A\sim\mu}\mathbb{E}_{H^t}\big[e^{\lambda\left(\hat{r}^{\mathrm{IS}}(A,H^t) - \overline{r}(A)\right)}\big] \leq e^{\frac{\lambda^2}{8t\varepsilon^2}}.$$

Applying first Fubini's theorem and then Donsker-Varadhan variational formula for $h(a) = \lambda\big(\hat{r}^{\mathrm{IS}}(a, h^t) - \overline{r}(a)\big)$, we get

$$\mathbb{E}_{H^t}\left[e^{\sup_{\pi\in\Pi}\mathbb{E}_{A\sim\pi}[\lambda\left(\hat{r}^{\mathrm{IS}}(A,H^t) - \overline{r}(A)\right)] - D_{\mathrm{KL}}(\pi \,\|\, \mu) - \frac{\lambda^2}{8t\varepsilon^2}}\right] \leq 1.$$

The end of the proof uses Chernoff's bound. For any $\alpha > 0$, with probability smaller than $e^{-\alpha}$ over H^t, we have

$$\sup_{\pi\in\Pi} \lambda\Big(\widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t) - \mathscr{R}(\pi)\Big) - D_{\mathrm{KL}}(\pi \,\|\, \mu) - \frac{\lambda^2}{8t\varepsilon^2} > \alpha.$$

As the statement holds for the supremum over $\pi \in \Pi$, it then holds for any $\pi \in \Pi$, even those depending on the observed data H^t. We set $\beta/2 = e^{-\alpha}$, that is $\alpha = \ln(2/\beta)$, and take the complement. Rearranging terms, we get that with probability no smaller than $1 - \beta/2$, for all $\pi \in \Pi$ simultaneously, we have

$$\widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t) - \mathscr{R}(\pi) \leq \frac{D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln(2/\beta)}{\lambda} + \frac{\lambda}{8t\varepsilon^2}.$$

Using the same argument to $\{n\big(\overline{r}(a) - \hat{r}^{\mathrm{IS}}(a, H^n)\big)\}_{n=1}^t$ and a union bound over the two results concludes the proof. ∎

To derive a tighter bound in ε, one can use Bernstein's inequality instead of Hoeffding-Azuma's inequality and obtain a *Bernstein PAC-Bayes bound* [4, Theorem 7].

Proposition 2 (Bernstein PAC-Bayes bound for $\hat{r}^{\mathrm{IS}}$): Let $\mu \in \Pi$ be any prior policy independent of H^t and let ε be a uniform lower bound on $\{\pi_n(a)\}_{n=1}^t$ for all $a \in \mathcal{A}$ (resp. on $\{\pi_n(a, x)\}_{n=1}^t$ in the contextual bandits setting) . Then, for any $\lambda \in (0, 1)$, and any $\beta \in (0, 1)$, with probability no smaller than $1 - \beta$ (over the sampling of H^t)

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| \leq \frac{\lambda 2(e-2)}{t\varepsilon} + \frac{D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln\frac{2}{\beta}}{\lambda},$$

holds simultaneously for all policy $\pi \in \Pi$ depending on H^t.

Proof: Let $a \in \mathcal{A}$ be a fixed action and let $Z_n^{\mathrm{IS}}(a)$ be as defined in eq. (6). We define $V_t^{\mathrm{IS}}(a) := \sum_{n=1}^t \mathbb{E}[(Z_n^{\mathrm{IS}}(a))^2 | Z_1^{\mathrm{IS}}(a), \ldots, Z_{n-1}^{\mathrm{IS}}(a)]$ the variance of the martingale $M_t^{\mathrm{IS}}(a)$ defined in eq. (5). The proof then follows similarly as the proof of Proposition 1 using Bernstein's inequality instead of Hoeffding-Azuma's inequality. The final result is obtained using [3, Lemma 2] to get the upper bound $\mathbb{E}_{A\sim\pi}[V_t^{\mathrm{IS}}(A)] \leq \frac{2t}{\varepsilon}$ on the expected variance of the martingale. ∎

C. Parameter-free PAC-Bayes bound for $\widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)$

Under their current forms, the bounds in Proposition 1 and Proposition 2 are not readily usable; we still need to choose a value of $\lambda > 0$. In Proposition 1, if we could select the parameter λ after observing H^t, we would choose the parameter that minimizes the bound that is we would use $\lambda = 2\varepsilon\sqrt{2t(D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln(2/\beta))}$. For this value, the bound in Proposition 1 becomes

$$|\widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t) - \mathscr{R}(\pi)| \leq \frac{1}{\varepsilon}\sqrt{\frac{D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln(2/\beta)}{2t}}.$$

Similarly, if we could set $\lambda = \sqrt{\frac{t\varepsilon(D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln(2/\beta))}{2(e-2)}}$ in Proposition 2, we would optimize the right-hand side and obtain a bound of

$$|\widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t) - \mathscr{R}(\pi)| \leq \sqrt{\frac{8(e-2)(D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln(2/\beta))}{t\varepsilon}}.$$

However, this is not possible since the parameter needs to be selected *before* the draw of H^t and can, therefore, not depend on the realization of this data [25, Remark 14].

One idea to circumvent this difficulty is to consider a grid of values $\{\lambda_i\}_{i=1}^K$ of λ such that the corresponding bound holds and with probability $\{1 - \beta\}_{i=1}^K$. Applying a union bound argument, one gets a bound holds for all $\{\lambda_i\}_{i=1}^K$ simultaneously with probability at least $1 - \sum_{i=1}^K \beta_i$. This technique is used in [17] but fails to attain a rate as tight as the one associated with the optimal value of λ. Our next results improve on the previous approach and provide *almost optimal* bounds achieving the same rate as the optimal bound.

Theorem 1 (Optimized Hoeffding PAC-Bayes bound): Let $\mu \in \Pi$ be any prior policy independent of H^t and let ε be a uniform lower bound on $\{\pi_n(a)\}_{n=1}^t$ for all $a \in \mathcal{A}$. Then, for any $\beta \in (0, 1)$, with probability no smaller than $1 - \beta$ (over the sampling of H^t), we have that

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| \leq \frac{1}{\varepsilon}\sqrt{\frac{D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln\frac{4\pi}{3\beta}}{t}},$$

holds simultaneously for all policy $\pi \in \Pi$ depending on H^t.

Proof: Our proof is based on the technique introduced by [5] and refined in [26]. We start with the parametric bound from Proposition 1. At a high level, the idea is to form a grid over the events' space (that is, the possible values for $D_{\mathrm{KL}}(\pi \,\|\, \mu)$) and find the best parameter conditioned on each event in that grid. To do so, we define the event $\mathcal{E}_1 = \{D_{\mathrm{KL}}(\pi \,\|\, \mu) \leq 1\}$ and for all $k \in \{2, \ldots\}$, we set the event $\mathcal{E}_k = \{k - 1 < D_{\mathrm{KL}}(\pi \,\|\, \mu) \leq k\}$. Conditioned on the event $\mathcal{E}_k$, there are two parameters that we can tune, $\lambda_k > 0$ and $\beta_k \in [0, 1]$. We then define the event $\mathcal{B}_{\lambda_k,\beta_k}$ as the set $\{H^t : |\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \frac{\lambda_k}{8t\varepsilon^2} + \frac{D_{\mathrm{KL}}(\pi \,\|\, \mu) + \ln\frac{2}{\beta_k}}{\lambda_k}\}$. Then given the event $\mathcal{E}_k$, with probability no more than $\mathbb{P}[\mathcal{B}_{\lambda_k,\beta_k} | \mathcal{E}_k]$, there exists a dataset H^t such that

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \frac{\lambda_k}{8t\varepsilon^2} + \frac{k + \ln\frac{2}{\beta_k}}{\lambda_k},$$

where we used the fact that $D_{\mathrm{KL}}(\pi \parallel \mu) \leq k$ given $\mathcal{E}_k$. As the right hand-side does not depend on the data H^t anymore, we can optimize over λ_k. We first set $\beta_k = \frac{6\beta}{\pi k^2}$ (note that π denotes here the mathematical constant pi and not the policy as on the left hand-side of the inequality) and then set the optimal value $\lambda_k = \sqrt{8t\varepsilon(k + \ln \frac{2\pi k^2}{6\beta}})$. We get that with probability no more than $\mathbb{P}[\mathcal{B}_{\lambda_k,\beta_k}|\mathcal{E}_k]$,

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \frac{1}{\varepsilon}\sqrt{\frac{k + \ln \frac{\pi k^2}{3\beta}}{2t}}.$$

The square root is a non-decreasing function, and given $\mathcal{E}_k$, we have $k < D_{\mathrm{KL}}(\pi \parallel \mu) + 1$. It therefore comes, given $\mathcal{E}_k$, with probability no more than $\mathbb{P}[\mathcal{B}_{\lambda_k,\beta_k}|\mathcal{E}_k]$,

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \sqrt{\frac{D_{\mathrm{KL}}(\pi \parallel \mu) + \ln \frac{e\pi(D_{\mathrm{KL}}(\pi \parallel \mu)+1)^2}{3\beta}}{2t\varepsilon^2}}.$$

Since $x + \ln(\frac{e\pi(1+x)^2}{3\beta})$ is a non-decreasing, concave, continuous function for all $x > 0$, it can be upper bounded by its envelope. That is $x + \ln(\frac{e\pi(1+x)^2}{3\beta}) \leq \inf_{a>0}\left\{\frac{a+3}{a+1}x + \ln(\frac{e\pi(a+1)^2}{3}) - \frac{2a}{a+1}\right\}$. Setting $a = 1$, we get $x + \ln(\frac{e\pi(1+x)^2}{3\beta}) \leq 2x + \ln(\frac{4\pi}{3\beta})$. Using this bound, we get that given $\mathcal{E}_k$, with probability no more than $\mathbb{P}[\mathcal{B}_{\lambda_k,\beta_k}|\mathcal{E}_k]$,

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \frac{1}{\varepsilon}\sqrt{\frac{D_{\mathrm{KL}}(\pi \parallel \mu) + \ln \frac{4\pi}{3\beta}}{t}}. \tag{7}$$

We now define $\mathcal{B}'$ as the event described in eq. (7). We have $\mathbb{P}[\mathcal{B}'|\mathcal{E}_k]\mathbb{P}[\mathcal{E}_k] \leq \mathbb{P}[\mathcal{B}_k|\mathcal{E}_k]\mathbb{P}[\mathcal{E}_k] \leq \mathbb{P}[\mathcal{B}_k] \leq \beta_k = \frac{6\beta}{\pi k^2}$. Therefore the probability of $\mathcal{B}'$ is bounded as $\mathbb{P}[\mathcal{B}'] = \sum_{k=1}^{\infty}\mathbb{P}[\mathcal{B}'|\mathcal{E}_k]\mathbb{P}[\mathcal{E}_k] \leq \sum_{k=1}^{\infty}\frac{6\beta}{\pi k^2} = \beta$. We then have $\mathbb{P}[\mathcal{B}'^C] = 1 - \mathbb{P}[\mathcal{B}'] \geq 1 - \beta$ which concludes our proof. ■

Theorem 2 (Optimized Bernstein PAC-Bayes bound): Let $\mu \in \Pi$ be any prior policy independent of H^t and let ε be a uniform lower bound on $\{\pi_n(a)\}_{n=1}^t$ for all $a \in \mathcal{A}$. Then for any $\beta \in (0, 1)$, with probability no smaller than $1 - \beta$ (over the sampling of H^t), we have that

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| \leq 2\sqrt{\frac{(e-2)(D_{\mathrm{KL}}(\pi \parallel \mu) + \ln \frac{4\pi}{3\beta})}{t\varepsilon}},$$

holds simultaneously for all policy $\pi \in \Pi$ such that $D_{\mathrm{KL}}(\pi\|\mu)$ is smaller than $\frac{2(e-2)-t\varepsilon(2+\ln\frac{2}{\beta})}{2t\varepsilon}$.

Proof: The proof follows by applying the same technique as for Theorem 1. As by assumption $D_{\mathrm{KL}}(\pi \parallel \mu) \leq \frac{2(e-2)-t\varepsilon(2+\ln\frac{2}{\beta})}{2t\varepsilon}$, we can define the event the events $\mathcal{E}_k = \{k - 1 < D_{\mathrm{KL}}(\pi \parallel \mu) \leq k\}$ only for $k \in \{1, \ldots, K\}$, where $K = \left\lceil\frac{2(e-2)-t\varepsilon(2+\ln\frac{\pi}{3\beta})}{2t\varepsilon}\right\rceil$. Then we set the event $\mathcal{B}_{\lambda_k,\beta_k}$ as the set $\{H^t : |\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \frac{\lambda_k 2(e-2)}{t\varepsilon} + \frac{D_{\mathrm{KL}}(\pi \parallel \mu)+\ln\frac{2}{\beta_k}}{\lambda_k}\}$. Then given the event $\mathcal{E}_k$, with probability no more than $\mathbb{P}[\mathcal{B}_{\lambda_k,\beta_k}|\mathcal{E}_k]$, there exists a dataset H^t such that

$$|\mathscr{R}(\pi) - \widehat{\mathscr{R}}^{\mathrm{IS}}(\pi, H^t)| > \frac{\lambda_k 2(e-2)}{t\varepsilon} + \frac{k + \ln\frac{2}{\beta_k}}{\lambda_k},$$

where we used the fact that $D_{\mathrm{KL}}(\pi \parallel \mu) \leq k$ given $\mathcal{E}_k$. As before, we set $\beta_k = \frac{6\beta}{\pi k^2}$ and optimize over λ_k, that is we set the optimal value $\lambda_k = \sqrt{\frac{t\varepsilon(k+\ln\frac{2\pi k^2}{6\beta})}{2(e-2)}}$. The value of λ_k is the largest for $k = K$, and we can verify that $\lambda_K = \sqrt{\frac{t\varepsilon(K+\ln\frac{2\pi K^2}{6\beta})}{2(e-2)}} \leq \sqrt{\frac{t\varepsilon(2K+1+\ln\frac{\pi}{3\beta})}{2(e-2)}} \leq 1$. The proof then concludes in a similar way as for Theorem 1. ■

IV. Conclusion

In this paper, we provide refined PAC-Bayes bounds for bandit problems, extending the analysis from [2]. We propose two parameter-free bounds, one based on Hoeffding-Azuma inequality, the other on the Bernstein inequality, that achieve optimal rates through a new "in probability" parameter optimization technique. Future research will focus on leveraging these refined reward estimate bounds to derive new PAC-Bayes regret bounds.

References

[1] Y. Seldin and N. Tishby, "PAC-Bayesian analysis of co-clustering and beyond." *Journal of Machine Learning Research*, vol. 11, no. 12, 2010.

[2] Y. Seldin, F. Laviolette, J. Shawe-Taylor, J. Peters, and P. Auer, "PAC-Bayesian Analysis of Martingales and Multiarmed Bandits," May 2011, arXiv:1105.2416 [cs]. [Online]. Available: http://arxiv.org/abs/1105.2416

[3] Y. Seldin, N. Cesa-Bianchi, P. Auer, F. Laviolette, and J. Shawe-Taylor, "PAC-Bayes-Bernstein Inequality for Martingales and its Application to Multiarmed Bandits."

[4] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-Bayesian inequalities for martingales," *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7086–7093, 2012.

[5] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, "More PAC-Bayes bounds: From bounded losses, to losses with general tail behaviors, to anytime validity," Jun. 2024, arXiv:2306.12214 [stat]. [Online]. Available: http://arxiv.org/abs/2306.12214

[6] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on Applications of Multi-Armed and Contextual Bandits," in *2020 IEEE Congress on Evolutionary Computation (CEC)*. Glasgow, United Kingdom: IEEE, Jul. 2020, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9185782/

[7] N. Silva, H. Werneck, T. Silva, A. C. Pereira, and L. Rocha, "Multi-Armed Bandits in Recommendation Systems: A survey of the state-of-the-art and future directions," *Expert Systems with Applications*, vol. 197, p. 116669, Jul. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417422001543

[8] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[9] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "A framework for structural risk minimisation," in *Conference on Computational learning theory (COLT)*, 1996, pp. 68–76.

[10] D. A. McAllester, "Some PAC-Bayesian theorems," in *Conference on Computational learning theory (COLT)*, 1998, pp. 230–234.

[11] ——, "PAC-Bayesian model averaging," in *Conference on Computational learning theory (COLT)*, 1999, pp. 164–170.

[12] ——, "PAC-Bayesian stochastic model selection," *Machine Learning*, vol. 51, no. 1, pp. 5–21, 2003.

[13] O. Catoni, "PAC-Bayesian supervised classification: The thermodynamics of statistical learning," *IMS Lecture Notes Monograph Series*, vol. 56, p. 163pp, 2007.

[14] J. Peters, K. Mulling, and Y. Altun, "Relative Entropy Policy Search," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, no. 1, pp. 1607–1612, Jul. 2010. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/7727

[15] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust Region Policy Optimization," Apr. 2017, arXiv:1502.05477 [cs]. [Online]. Available: http://arxiv.org/abs/1502.05477

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Aug. 2017, arXiv:1707.06347 [cs]. [Online]. Available: http://arxiv.org/abs/1707.06347

[17] Y. Seldin, F. Laviolette, N. Cesa-Bianchi, J. Shawe-Taylor, and P. Auer, "PAC-Bayesian Inequalities for Martingales," Jul. 2012, arXiv:1110.6886 [cs]. [Online]. Available: http://arxiv.org/abs/1110.6886

[18] A. Swaminathan and T. Joachims, "Counterfactual Risk Minimization," in *Proceedings of the 24th International Conference on World Wide Web*. Florence Italy: ACM, May 2015, pp. 939–941. [Online]. Available: https://dl.acm.org/doi/10.1145/2740908.2742564

[19] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, ser. Adaptive computation and machine learning. Cambridge, Mass: MIT Press, 1998.

[20] H. Flynn, D. Reeb, M. Kandemir, and J. Peters, "PAC-Bayes Bounds for Bandit Problems: A Survey and Experimental Comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 15 308–15 327, Dec. 2023, arXiv:2211.16110 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2211.16110

[21] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain Markov process expectations for large time, I," *Communications on Pure and Applied Mathematics*, vol. 28, no. 1, pp. 1–47, 1975.

[22] W. Hoeffding, "PROBABILITY INEQUALITIES FOR SUMS OF BOUNDED RANDOM VARIABLESl."

[23] K. Azuma, "Weighted Sums of Certain Dependent Random Variables," *Tohoku Mathematical Journal, Second Series*, vol. 19, no. 3, pp. 357–367, 1967.

[24] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[25] P. K. Banerjee and G. Montúfar, "Information complexity and generalization bounds," in *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021, pp. 676–681.

[26] B. Rodríguez-Gálvez, R. Thobaben, and M. Skoglund, "An Information-Theoretic Approach to Generalization Theory," Aug. 2024, arXiv:2408.13275 [stat]. [Online]. Available: http://arxiv.org/abs/2408.13275